This syllabus is a living and breathing document. It is likely to change throughout the course. If and when it does, I will alert to such changes via NYU Classes.

DS-GA-1001: Introduction to Data Science

Thursday Evenings

Meyer Hall

Professor: Brian d'Alessandro
Email: bdalessa@stern.nyu.edu
Office Location: It varies
Office Hours: By appointment

Section Leader: Leslie Huang
Email: lesliehuang@nyu.edu
Office Location: CDS
Office Hours: Wed 4 - 6 pm

Section Leader: Lee Tanenbaum
Email: leedtan@gmail.com
Office Location: CDS
Office Hours: Tues 5:30 - 7:30 pm

1 Course Description

Businesses, governments, and individuals create massive collections of data as a by-product of their activity. Increasingly, decision-makers and systems rely on intelligent technology to analyze data systematically in order to improve decision-making. In many cases automating analytical and decision-making processes is necessary because of the volume of data and the speed with which new data are generated.

We will examine how data analysis technologies can be used to improve decision-making. We will study the fundamental principles and techniques of data science, and we will examine real-world examples and cases to place data science techniques in context, to develop data-analytic thinking, and to illustrate that proper application is as much an art as it is a science. In addition, we will work hands-on with the Python programming language and its associated data analysis libraries.

After taking this course you should be able to:

1. Formulate a solution strategy to data science problems using the complete data mining process, including problem formulation, exploratory analysis, modeling, evaluation, implementation, and feedback
2. Apply basic exploratory analysis to identify abnormalities in data (i.e., missing values, outliers, redundant features, etc.)
3. Anticipate and identify ways in which sampled data may be biased

4. Prepare a data set for supervised learning, including constructing labels and appropriately splitting the data for training, validation and testing
5. Identify instances of data leakage and apply techniques to avoid it
6. Perform the appropriate feature transformations for processing categorical data and for making non-linear representations in linear models
7. Identify the appropriate set of algorithms (i.e., regression vs. decision tree vs. clustering) for a given problem statement, and give an appropriate analysis of the pros/cons of each for the problem at hand
8. Identify the appropriate evaluation metric based on a given problem goal (i.e., AUC vs. MSE)
9. Use Python's Scikit-learn package to solve classic classification/regression problems
10. Find an optimal model fit by tuning hyperparameters on a variety of classification algorithms
11. Explain model regularization, both mathematically and at the level appropriate for a smart lay-person, including the pros/cons of L1 vs L2 regularization.
12. Explain the underlying mathematics of matrix factorization, and identify multiple appropriate uses for matrix factorization techniques in data mining problems
13. Set up and describe recommender system problems in both an unsupervised or supervised context
14. Perform model selection using cross-validation
15. Design and implement a decision function that reflects the cost of classification errors that might be made
16. Discuss ethical implications surrounding privacy, data sharing, and algorithmic decision-making for a given data science solution

2 Focus and Interaction

The course will explain through lectures and real-world examples the fundamental principles, uses, and appropriate technical details of machine learning, data mining and data science. The emphasis is primarily on understanding the fundamental concepts and applications of data science. We will cover several algorithms though this is not an algorithms course, nor a course in machine learning or computational theory. Our aim rather is to present fundamental algorithms within the context of a larger data science and decision making process.

I will expect you to be prepared for class discussions by having satisfied yourself that you understand what we have done in the prior classes. The assigned readings will cover the fundamental material. The class meetings will be a combination of lectures/discussions on the fundamental material, discussions of business applications of the ideas and techniques, case discussions, student exercises, and demos.

You are expected to attend every class session, to arrive prior to the starting time, to remain for the entire class, and to follow basic classroom etiquette, including (unless otherwise directed) using electronic devices except where necessary to follow along with the lecture or lab.
In general, we will follow NYU default policies unless I state otherwise. I will assume that you have read them and agree to abide by them:

http://gsas.nyu.edu/page/policiesprocedures

The NYU Classes site for this course will contain lecture notes, reading materials, assignments, and late-breaking news.

If you have questions about class material that you do not want to ask in class, or that would take us well off topic, please detain me after class, come to office hours to see me or the TAs, or ask on the discussion board. The discussion board is the preferred method of asking questions, as others may benefit from the answers being available on NYU Classes. As a corollary to this, please try to answer your classmates questions.

3 Readings/Texts

This is a graduate course so we‘ll assume that you have the self-motivation and discipline to keep up with the readings on your own. The syllabus will note exactly what readings are required and which are optional (but still worth reading). We‘ll assign quizzes to test your understanding and absorption of the assigned, required material.

Required Reading

Title: *Data Science for Business*
Author(s)/Pub: Provost and Fawcett; O'Reilly 2013
Source: Online or in the NYU Book store

Title: *An Introduction to Statistical Learning*
Author(s)/Pub: James, Whitten, Hastie, Tibshirani; Springer 2013
Source: Free pdf at `https://www-bcf.usc.edu/~gareth/ISL/ISLR%20First%20Printing.pdf`

Additionally, I have crafted a series of iPython notebooks to supplement the above books on various technical details (regarding Python and math).

http://nbviewer.ipython.org/github/briandalessandro/DataScienceCourse/tree/master/ipython

or

https://github.com/briandalessandro/DataScienceCourse/tree/master/ipython

Supplemental Reading

The following books are not mandatory but are worth owning and mastering as part of your development as a data scientist. These particularly focus on doing data science with Python.

Title: *Python for Data Analysis*
Author(s)/Pub: McKinney; O'Reilly 2012
Source: Online or in the NYU Library

Title: *Data Science from Scratch*
Author(s)/Pub: Grus; O'Reilly 2015
Source: Online or in the NYU Library

Additional Considerations

Readings: A lot of valuable reference material is not published as a text book but instead comes in the form of academic and conference white papers. Throughout this course, we‘ll review several papers as mandatory readings. We‘ll also post several more as recommended supplementary material. In general, learning to read through these types of papers will only help you in a career that requires continuous learning.

Lecture notes: In an effort to conserve resources, lecture materials will be handed out digitally. I expect you to ask questions about any material in the notes that is unclear after our class discussion and reading the book. Having the book frees up class time for more discussion of applications, cases, etc.-many of your questions may be answered in the book. If any are not, please let me know! Depending on the direction our class discussion takes, we may not cover all material in the class notes for any particular session. If the notes and the book are not adequate to explain a topic we skip, you should ask about it on the discussion board. I will be happy to follow up.

Please don't hesitate to come and talk to me about what supplemental material might be best for you, if you want to go further on any topics covered in the course.

4 Requirements and Grading

Grading Rubric:

Homework	30%
Term Project	25%
Quizzes	20%
Final Exam	25%

At the Center for Data Science we seek to teach challenging courses that allow students to demonstrate differential mastery of the subject matter. Assigning grades that reward excellence and reflect differences in performance is important to ensuring the integrity of our curriculum. In my experience, students generally become engaged with this course and do excellent or very good work, receiving A's and B's, and only one or two perform only adequately or below and receive Cs or lower. Note that the actual distribution for this course and your own grade will depend upon how well each of you actually perform this particular semester.

4.1 Homework Assignments

The homework assignments are listed (by assignment date) in the class schedule below. This is the authoratitive source on assignments and due dates. However, from time to time, and for various reasons, exceptions may occur. In such a case the syllabus will be updated and the changes will be announced both in class and on NYU Classes. Except as explicitly noted (see next paragraph), you are expected to complete your assignments on your own-without interacting with other students on the completion of your assignment. You are free of course to discuss the concepts with your classmates, and to discuss similar problems to the ones in the homework.

I hope with the support of myself, the TAs, and your classmates, we operate under a "diligent attempt but limited frustration“ policy: (1) If you get stuck on something, spend some time Googling to try to find the answer. If you seem to be moving forward, keep going. That search and discovery will pay off, both in terms of the direct learning about how to do what you need to do, and also in terms of your learning how to find such things out. (E.g., if you dont know what Stackoverflow is, you will learn!). BUT, (2) limit frustration-start your assignments early enough that if you run into a wall, you can just stop searching and ask about it. Lets say, if you feel like you have not moved forward after 15 minutes of being stuck, just stop and ask: your classmates, on the discussion board, to the TAs. If you dont get a solution, escalate it to me.

Completed assignments must be handed on NYU Classes at least one hour prior to the start of class on the due date (that is, by 5pm), unless otherwise indicated. Assignments will be graded and returned promptly. Answers to homework questions should be well thought out and communicated precisely, as if reporting to your boss, client, or potential funding source. Avoid sloppy language, poor diagrams, irrelevant discussion, and irrelevant program output.

The hands-on tasks in the homework will be based on data that we will provide. You will mine the data to get hands-on experience in formulating problems and using the various techniques discussed in class. You will use these data to build and evaluate predictive models.

4.2 Prerequisites

You should be familliar with probability theory, linear algebra, statistics and multi-variate calculus. We will not teach these subjects in class, but knowledge is assumed in the course readings and some of the lecture material.

Some proficiency in computer programming is also required. Students with no programming experience at all, or little knowledge of Python programming are strongly encouraged to take Programming for Data Science along with this course. Without basic competency in Python programming you will likely struggle with the homework and class project.

4.3 Hands on Programming

Data Science is not possible without some sort of programming knowledge. This class will involve hands-on assignments and demonstrations of data science techniques with Python and its associated libraries. We will cover the installation of the iPython Notebook (http://ipython.org/notebook.html).

For those students with little to no programming experience (Python or not), it is highly advised to install iPython and start learning the language and platform in advance of the class. There are many online tutorials for getting started with the language.

IMPORTANT: You must have access to a computer on which you can install software. If you do not have such a computer, please see me immediately so we can make alernative arrangements. During class we will have live demos of certain analysis techniques in action, using the iPython programming environment.

Generally the section leaders should be the first point of contact for questions about and issues with the homeworks. If they cannot help you to your satisfaction, please do not hesitate to come see me.

4.4 Late Assignments

As stated above, assignments are to be submitted on NYU Classes at least one hour prior to the start of the class on the due date. Assignments up to 24 hours late will have their grade reduced by 25%; assignments up to one week late will have their grade reduced by 50%. After one week, late assignments will receive no credit. Please turn in your assignment early if there is any uncertainty about your ability to turn it in on time.

4.5 Term Project

A term project report will be prepared by student teams. We will give you the instructions on how to form your teams. Teams are encouraged to interact with the instructor and section leader electronically or face-to-face in developing their project reports. You will submit various milestone deliverables through the course. We will discuss the project requirements in class.

4.6 Final Exam

The final exam will be given on the last scheduled class date. The assignment will be administered online via NYU Classes. It is imperative that you bring a laptop to class on that day. The exam will be open book and open notes.

4.7 Regrading

If you feel that a calculation, factual, or judgment error has been made in the grading of an assignment or exam, please write a formal memo to me describing the error, within one week after the class date on which that assignment was returned. Include documentation (e.g., pages in the book, a copy of class notes, etc.). I will make a decision and get back to you as soon as I can. Please remember that grading any assignment requires the grader to make many judgments as to how well you have answered the question. Inevitably, some of these go “in your favor” and possibly some go against. In fairness to all students, the entire assignment or exam will be regraded.

FOR STUDENTS WITH DISABILITIES: If you have a qualified disability and will require academic accommodation during this course, please contact the Moses Center for Students with Disabilities (CSD, 998-4980) and provide me with a letter from them verifying your registration and outlining the accommodations they recommend. If you will need to take an exam at the CSD, you must submit a completed Exam Accommodations Form to them at least one week prior to the scheduled exam time to be guaranteed accommodation.

5 Course Outline

5.1 Week 1: 9/6/2018

Theme: What is Data Science?

5.1.1 Learning Objective(s)

1) Understand the skill set required to become a data scientist. 2) Formulate unstructured domain language problems into structured data science problems. 3) Articulate the the complete data mining process from end to end.

5.1.2 Readings

Required:

- DSforBus Ch 1 - 2
- iPython notebooks (NumPyBasics, SimpleiPythonExample).

Recommended:

- DsfromScratch Ch 1 - 2
- PDA Ch 1 - 2
- Analyzing the Analyzers (PDF NYU Classes)

5.1.3 Assignments

- *Assigned:* Homework 1 (due 9/14/2016)
- *Due:* None

5.1.4 Lab

Note: Bring a laptop!. We'll focus on setting up our Python environment, working with Github and running our first iPython notebooks.

5.2 Week 2: 9/13/2018

Theme: Data Analytic Thinking I

5.2.1 Learning Objective(s)

1). Translate unstructured domain problems into structured data science problems. 2). Properly sample data and expore it. 3). Think through common problem constraints.

5.2.2 Readings

Required:

- DSforBus Ch 3
- ISLR Ch 1, sec. 2.1
- iPython Notebook (PandasIntro)

Recommended:

- DsfromScratch Ch 3, 9-10
- PDA Ch 4
- Scientific Computing Best Practices (PDF in NYU Classes)

5.2.3 Assignments

- *Assigned:* None
- *Due:* Homework 1 (due 9/14/3016)

5.2.4 Lab

Note: Bring a laptop!. We'll cover some Numpy basics, write a logisitic regression scoring function and compare Numpy's efficiency against naive implementations.

5.3 Week 3: 9/20/2018

Theme: Informed Partitions

5.3.1 Learning Objective(s)

1). Apply recursive partitioning to classify instances. 2). Identify the different hyper-parameter dimensions Decision Trees and know how to tune them.

5.3.2 Readings

Required:

- DSforBus Ch 4
- ISLR sec. 3.1-3.3,8.1
- iPython notebooks (DecisionTrees)

Recommended:

- DsfromScratch Ch 17
- PDA Ch 5
- Primer on Information Theory Bishop Ch 1.6 (PDF NYU Classes)

5.3.3 Assignments

- *Assigned:* Homework 2 (due 10/4/2018)
- *Due:* None

5.3.4 Lab

Note: Bring a laptop!. We'll do some data exploration and visualization using Pandas and Matplotlib.

5.4 Week 4: 9/27/2018

Theme: Fitting a Mathematical Model to Data

5.4.1 Learning Objective(s)

1). Understand how mathematical models are fit to data. 2). Understand loss functions and the principles of Expected Risk Minimization. 3). Learn the mechanics of linear separating hyperplanes and how to fit and tune them.

5.4.2 Readings

Required:

- DSforBus Ch 5
- ISLR sections 4.1-4.3, 9.1-9.3
- iPython Notebooks (ERM_LogReg, SVM)

Recommended:

- DSfromScratch Ch 16
- PDA Ch 6
- A User's Guide to Support Vector Machines (PDF NYU Classes)
- Making Classifiers Robust to Sample Selection Bias (PDF NYU Classes)

5.4.3 Assignments

- *Assigned:* None
- *Due:* None

5.4.4 Lab

Note: Bring a laptop!. Churn Analysis Case Study part 1.

5.5 Week 5: 10/4/2018

Theme: Overfitting and its Avoidance

5.5.1 Learning Objective(s)

1). Design a predictive modeling experiment. 2). Know why all models are wrong, though some are useful. 3). Understand the drivers of bias and variance in model estimation error.

5.5.2 Readings

Required:

- DSforBus Ch 7, 8
- ISLR sections 2.2, 5.1-5.2
- iPython Notebooks (SimpleOverfittingExample, Resampling, BiasVariance)

Recommended:

- DSfromScratch Ch 11
- PDA Ch 7, 8
- A Few Useful Things to Know About Machine Learning

5.5.3 Assignments

- *Assigned:* Homework 3 (due 10/18/2018)
- *Due:* Homework 2

5.5.4 Lab

Note: Bring a laptop!. TBD

5.6 Week 6: 10/11/2018

Theme: The Art and Science of Evaluation

5.6.1 Learning Objective(s)

1). Determining which evaluation metrics are appropriate for a given task. 2). How to apply the expected value framework to decision making. 3). Work through cost-sensitive learning to make more optimal decisions

5.6.2 Readings

Required:

- iPython Notebook (Metrics_Ranking)

Recommended:

- https://en.wikipedia.org/wiki/Receiver_operating_characteristic
- PDA Ch 9

5.6.3 Assignments

- *Assigned* None
- *Due* None

5.6.4 Lab

Note: Bring a laptop!.

5.7 Week 7: 10/18/2018

Theme: Making Model Features and Discarding Them.

5.7.1 Learning Objective(s)

1). Engineer features to capture non-linear effects in linear models. 2). Apply feature selection to high-dimensional problems. 3). Understand the mathematics of regularization and how to tune a model using this method.

5.7.2 Readings

Required:

- DSforBus Ch 11, 12
- ISLR 6.1-6.3
- iPython Notebook (Regularization, FeatureSelection, Binning_NonLinear)

Recommended:

- DSfromScratch ch 9, 10
- PDA Ch 12

5.7.3 Assignments

- *Assigned:* Homework 4 (Due 11/2/16)
- *Due:* Homework 3

5.7.4 Lab

Note: Bring a laptop!.

5.8 Week 8: 10/25/2018

Theme: Using Neighbors for Supervised and Unsupervised Learning

5.8.1 Learning Objective(s)

1). Identify and formulate unsupervised learning problems appropriately. 2). Build clusters using both k-Means and hierarchical methods. 3). Evaluate and explain the results of clustering methods. 4). Use distance metrics for classification.

5.8.2 Readings

Required:

- DSforBus Ch 6
- ISLR section 10.1-10.2
- iPython Notebook (Clustering)

Recommended:

- DSfromScratch ch 12, 19

5.8.3 Assignments

- *Assigned:* None
- *Due:* None

5.8.4 Lab

Note: Bring a laptop!.

5.9 Week 9: 11/1/2018

Theme: Data Mining Through the Lens of a Spam Filter

5.9.1 Learning Objective(s)

1). Think through design considerations of a production quality classification system. 2). Learn and apply Naive Bayes for classification. 3). Create a numeric design matrix out of text data.

5.9.2 Readings

Required:

- DSforBus Ch 9, 10
- iPython Notebook (TextMining)

Recommended:

- DSfromScratch Ch 13, 20

5.9.3 Assignments

- *Assigned:* Homework 5 (Due 11/30/16)
- *Due:* Homework 4

5.9.4 Lab

Note: Bring a laptop!.

5.10 Week 10: 11/8/2018

Theme: Combining Models to Make Better Models

5.10.1 Learning Objective(s)

1). Identify situations in which combing models may lead to better performance. 2). Understand the mechanics and apply Random Forests and Gradient Boosting to solve supervised learning problems.

5.10.2 Readings

Required:

- DSforBus Ch 13, 14
- ISLR section 8.2
- iPython Notebooks (Bagging_RandomForests, GradientTreeBoosting)

Recommended:

- None

5.10.3 Assignments

- *Assigned:* None
- *Due:* None

5.10.4 Lab

Note: Bring a laptop!.

5.11 Week 11: 11/15/2018

Theme: Recommendations Through Collective Intelligence

5.11.1 Learning Objective(s)

1). Articulate recommendation problems as both supervised and unsupervised learning 2). Set up appropriate data for recommendation problems 3). Evaluate recommendation problems, but offline and online.

5.11.2 Readings

Required:

- None
- None

Recommended:

- White Paper Handout

5.11.3 Assignments

- *Assigned:* None
- *Due:* None

5.11.4 Lab

Note: Bring a laptop!.

5.12 Week 12: 11/22/2018

Thanksgiving recess. No class!

5.13 Week 13: 11/29/2018

TBD

5.14 Week 14: 12/6/2018

TBD

5.15 Week 15: 12/13/2018

Final Exam.